Published 2000 in Evangelos Dermatas (ed.), *Proceedings of COMLEX2000: Computational lexicography* (Patras: Patras University Press), 141-44.

http://www.sophia.de

HOW DO WE READ A DICTIONARY (AS MACHINES AND AS HUMANS)?

KINDS OF INFORMATION IN DICTIONARIES, CONSTRUCTED AND RECONSTRUCTED

Vincent C. Müller
American College of Thessaloniki
PO Box 21021, 55510 Pylea, Greece
vmueller@ac.anatolia.edu.gr
http://www.anatolia.edu.gr/act/

ABSTRACT

Two large lexicological projects for the *Center for the Greek Language*, Thessaloniki, were to be published in print and on the WWW, which meant that two conversions were needed: a near-database file had to be converted to fully formatted file for printing and a fully formatted file had to be converted to a database for WWW access. As it turned out, both conversions could make use of existing clues that indicated the kinds of information contained in each particular piece of text, thus separating fields from each other and ordering them into a tree-like structure. This indicates that both forms of the dictionaries, print and database, stem from the same cognitive need to categorize information into a *kind of information* before further understanding – be this for a human reader or for a machine.

Keywords: conversion, kinds, kinds of information, tagging, cognition, tree-structure, Visual Basic, WWW

1. INTRODUCTION – THE TASK

The theoretical issue to be discussed in this paper presented itself in the context of two large-scale dictionary projects of the *Center for the Greek Language*, Thessaloniki (*Kéntro Ellinikís Glóssas*, http://www.komvos.edu.gr), sponsored by the Greek Ministry for Education. Both dictionaries were to be published in print and on the WWW.

The two dictionaries concerned are: (1) A dictionary of Medieval Demotic Greek (1100–1669) with explanations in Modern Greek, edited by Emmanuél Kriarás – commonly known as the "Kriaras Dictionary". The publication of the first volumes started 30 years ago and now reaches volume 15, covering up to "Pi". Given the size and the need to update and homogenize the existing volumes, a concise version of the dictionary was prepared under the auspices of the *Center*, the first volume of which would cover "Alpha-Kappa". This concise dictionary was to be prepared for print and WWW accessible Oracle database. (2) The "Georgakas", Modern Greek-English dictionary, which was started 40 years ago in the US, accumulating a large corpus (analyzed on ca. 2 million cards) and attempting comprehensive coverage of Modern Greek. During the lifetime of the original editor, volume one, "Alpha", was compiled. This volume was re-edited and corrected in the *Center for the Greek Language* to be prepared for print (around 3.000 pp.) and WWW accessible Oracle database.

Thus, for each project, two kinds of files were needed: a fully formatted file for print publication and a set of database tables in ASCII that could be read into the Oracle relational database system. The latter also required all characters to be HTML compliant, i. e. characters beyond the basic first 128 of the standard web fonts (be these Roman or Greek) had to be given via their ampersand-codes or UTF-8 numbers. The lexicological basis was provided by the two philological teams, while retrieval from the database system and web-interface were designed by a separate team (for general problems, see [1]). However, in the two projects, only of one of the two needed kinds of files existed: In the case of Kriaras, we had a set of tagged ASCII files plus a routine to convert these into database tables. Philological corrections were made on these files but we had no files for print. In the case of Georgakas, we had fully formatted "Word 98 for Macintosh" files, into which philological corrections were made but no files for the database.

The project outlined here thus involved two tasks: (a) Convert the near-database files for Kriaras into fully typeset files. (b) Convert the typeset Georgakas files into database tables. Seemingly: Convert the computer readable form into the human readable form and inversely (cf. [2] and [4]).

2. PROCEDURES

To get to the theoretical point about the cognitive importance of categorization into kinds of things, let us have a brief look at the procedure for Kriaras. A program was written in Visual Basic that would use the existing tags and apply appropriate formatting to the respective "field", e. g. an piece of information tagged as "meaning, level 2" would get the appropriate numbering (a capital Greek letter), be formatted in the appropriate character style and followed by a period plus space. You can see that what divides this field from the previous one will depend on what kind of field *that previous one* is (e. g. a "meaning level 1" or "meaning level 2", a bibliographic reference, …). The same applies for the separator to be inserted afterwards – e. g. if a bibliographical reference follows, you should not put a period, but just a space. So the program needs to "keep in mind" where in a particular entry this field is situated (we need to keep that in mind too, it is crucial for the theoretical point). Technically, the program kept "switches" on/off, to record what kind of fields had already occurred in a lemma, on which field the particular field in question depended (for the notion of "dependence", see below). This program grew ever more complicated as in each run of corrections, the philologists discovered more rules for special cases, "yes, this is followed by that separator unless this other kind comes first, but if that special thing comes afterwards, then, of course, we need that other separator…". Frequently, it was discovered that rules needed to be invented for standardization – rules that could be bent or left unspecified in previous practice for print now had to be defined for the program. What was it that we were doing here, for the printed version? We were separating the kinds of information in such a way as to make them transparent to the human reader of the dictionary; also making clear which piece of information depends on which other piece of information.

This become clearer in the other task: converting the formatted Georgakas files into fields – a task that initially seemed harder, but actually turned out to be easier (which shows that it was not a case of Natural Language Processing). We had a predefined set of fields for the database, which was altered only slightly. The program could use existing formatting information to find where a field begins, where it ends and what *kind* of field it is. More importantly, it could use information about sequence: the beginning of a paragraph is the headword itself, never the etymology, etc. So, what comes at the beginning of a paragraph in bold is always the first field (if it contains a space plus hyphen a special routine must be run), ending where the bold print ends; the comma needs to be removed. Etymology is at the end in angular brackets, meanings of level 2 have a Roman numeral of a particular font, etc. etc. Some rules will be a little more complicated, like: "what comes after the end of bold print, is in Roman characters (not Greek), not italics, neither in round nor in angular brackets, *this* is the only meaning of the article, unless there is a number in that particular font there, or the phrase 'see …'". Given a set of fields that was not excessively fine-grained, this task could be fulfilled with a small degree of error. How is that? "Reading" the dictionary, the computer just needed to re-construct what kinds of information the lexicographers had inserted into a particular entry. To facilitate usage for the human readers, they had consistently separated that information in more or less non-ambiguous ways and the computer could pick up on these separations. – The very same separations we had to insert into the printed version of Kriaras.

3. THEORETICAL BASIS: KINDS OF INFORMATION

Let us step back from the computer reading the Georgakas dictionary (or making the Kriaras dictionary human readable) and look at what a human does, when reading a dictionary. The fact that the tasks described above could actually be performed gives us a hint at the cognitive processes used by both humans and machine in the reading of the dictionary. It also reminds us of why good dictionaries look the way they do.

Imagine, someone uttering in your presence what sounds like "June". What are you to make of this utterance – is the person telling you her name, informing you about the month in which she was born, or what? In order to understand, you need context, but the issue here is not just that of the well known context-dependence of meaning and thus linguistic understanding, it is that you need the *kind of information* in order to understand. You need to know into which kind of information the

utterance falls in order to process it correctly (a name or a month). Given this fact, you will normally receive clues about that kind to help you understand. Just the same happens in a dictionary, too.

Imagine you are reading an entry in a dictionary, saying "**ílios**, o. sun." You need to know that the first expression is the Greek word you wanted to look up, followed by a piece of grammatical information (the article), followed by a translation into a specific language, English in this case. In other words, your information processing first needs to know the *kind* of the information to be processed – which is something that applies irrespective of whether the reader is a human or a machine. Traditionally, this tends to be overlooked when we say things like "only people can find senses, machines are better at finding clues" [3]. This is not quite right: people need clues, too. First, we have a clue what kind of (lexical) information we are looking at, *then* we can process its content.

We are given a chance to understand the kind of information by the help of clues a well-constructed dictionary in print will give us. The lexicographers will not just list the information, they will divide the kinds of information in a way that gives sufficient clues to the reader: print the headword in bold, divide it from the following by a comma, followed by the article (one out of a small finite list) which marks the headword as a noun, followed by a period, marking the end of the entry, the beginning of the explanations. We know that the next thing to expect is the explanation in English, which is also marked by a change from Greek to Roman font [not apparent in the above transliteration]. Further kinds of information (meaning levels, etymology, regional specification, etc.) will be marked with their specific clues. Imagine a dictionary entry devoid of any such clues; it would be quite unusable.

So, like ordinary mortals, lexicographers implicitly knew about kinds of information all along. In the process of constructing a dictionary, they imagined the "ideal" article and thought which kinds of information this would contain (which "fields" in database talk), how all desired information could be included. This complex grid of kinds had to be designed in the process of making a dictionary and then had to be made apparent to users in the traditional printed form. What is more, each piece of information is characterized by its kind plus its dependence in a "child-parent" relationship (I am indebted to Th. Kehagias at this point). This is most obvious in meaning levels, where each meaning of, say, level 3 is dependent on a specific meaning on level 2, not just on meaning level 2 in general. A meaning level 3 may at the same time be the parent of another meaning or (in the Kriaras dictionary) of a bibliographical reference. Each meaning level 2 is dependent on a meaning level 1, which is dependent on the headword. From the headword, some kinds depend directly, such as grammatical information or etymology. We are looking at a tree-structure, where each individual piece of information is uniquely characterized by a) the kind of information it is, and b) its parent. These two characteristics are interdefinable, given sequence: if you know what kind of information a piece belongs to and where it occurs in the sequence, you also know its parents, children and siblings. If you know the parent, know of what kind that is, and the sequence, you also know what kind of information this is – provided that a certain set of kinds is always used (e. g. there is no article without etymology).

So, when reading the dictionary, you know the pieces you read already, thus know your position in the tree and the possible branches that are available, thus knowing that only certain kinds of branches can occur when you get a clue (e.g. a bold Arabic number indicating meaning level 3) that tells you what kind of information the next piece will contain before you actually read and understand it. Given the tree structure, clues will depend on the parent/superior branch and on the preceding piece of information. The same kind will not always get the same clue: a meaning only gets a number if it is not the only meaning; several pieces of information of the same kind will be separated from each other in a way different from their separations towards other kinds – but this difference serves as a clue, too: don't expect a new kind now, the next piece is of the same kind.

This "inconsistency" of different clues before the same kind of information made the insertion of clues in the Kriaras dictionary somewhat difficult – but a rule for the insertion could be found in all cases (or created if no explicit rule existed). In Kriaras, tags were replaced by typographical clues while in the Georgakas dictionary typographical clues provided by the lexicographers had to be replaced by explicit tags that indicate fields. These tags were then used for the conversion to database tables. Actually, both tasks depended on the existence of clues in the respective existing forms and the insertion of different clues in the new forms.

Both forms of presenting the information are readable by both readers, machine and human - only that cognitive differences make each prefer different clues – but clues they both need in just the same way. In this sense, any properly constructed dictionary is readable by a machine.

3. WHAT IS A KIND?

The programs written for both conversion tasks just had to know which kinds of information existed and which clues there were (and were to be added). A traditional lexicographer, however, would not regard the printed version as a *form* of presenting the dictionary, for him/her, this *is* the dictionary. We can see now that this is to be deceived by appearances since both forms do essentially the same thing: present the lexical information in a tree-structure with clues as to the kind of information each piece belongs to.

Emphasizing the importance of categorization of lexical information into kinds leads to the questions, where these "kinds" come from and how are they related to each other. We shall only make some gestures towards these issues. First, some of these kinds fall into specific upper level categories of kinds of kinds, for instance, some kinds are "levels of meaning". Some can only occur dependent on other kinds, only as branches of a particular kind of branch or trunk (as children of particular parents, in the other common metaphor). Presumably, "kind A depends on kind B" just means "if kind B does not exist, then kind A does not exist either" or, what is logically equivalent, "if kind A exists, then kind B exists as well" – while the inverse is not true, of course. Dependence appears to be a fairly simple logical relation, in this case.

Furthermore, are these kinds "natural", i. e. existing independently of our knowledge of them, discovered by research and having boundaries that we should find? Or are they "nominal kinds" made by us, for our purposes, delineated at will? (Cf. [5].) The answer may not be so easy if one remembers that the second option implies relativity in the sense that one could make the kinds in any different way and be equally right. What would lexicographers' disputes be about, in this case? Just a fight over words, or should we not say "just" here? After all, ours is a fight over and with words, only that we now are reminded of their kinds again. In the design of a dictionary, we need to be very aware of the kinds of information to allow and the dependence relations they can have, whoever the envisaged reader may be.

4. REFERENCES

[1] Carr, M. (1997) "Internet Dictionaries and Lexicography" *International Journal of Lexicography* 10: (3) 209-230

[2] Fontenelle, T. (1997) *Turning a Bilingual Dictionary into a Lexical-Semantic Database* (Lexicographica Series Maior 79). Tübingen: M. Niemeyer

[3] Gazdar, G. (1999) "Word Senses and Computational Lexicography", (lecture 16) web page, updated 25.03.99. http://www.cogs.susx.ac.uk/lab/nlp/gazdar/teach/nlp/nlpnode144.html

[4] Heid, U. (1997) *Zur Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern* (Lexicographica Series Maior 77). Tübingen: M. Niemeyer

[5] Müller, V. C. (1999) *Realismus und Referenz: Arten von Arten*, PhD-thesis, Hamburg University